ForumAcusticum
2 0 0 5

From Learning Music to Learning to Separate

Paris Smaragdis
Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge MA 02139, USA, e-mail: **paris@merl.com**

Making machines that can understand musical structure has long been one of the holy grails of audio processing, separating overlapping sounds has been another. Here we present a simple framework initialy used for the first task, which has come to make itself very useful for source separation. We show that the same type of reasoning that allows one to find the building elements of a musical audio stream can also be used to find and extract elements contained in auditory scenes. We relate this work with recent developments in sparse representations and dimensionality reduction and show its application in a variety of situations.

1 Introduction

Research in musical signal analysis and source separation has historically been technically and philosophically unrelated. Musical signal analysis often involves trying to estimate how traditional musical constructs can explain a musical signal, whereas source separation has been a field using varied approaches to separate sounds. Although most often connections between these two areas are not drawn, recent research has been exposing a converging trend. Source separation as a field entered a major phase with the advent of sparse coding and representations. This was a natural evolution from the earlier ICA work (excellently surveyed in [5]) which introduced some of the first successful statistical models for source separation. Using the same statistical models researchers soon found out that representing signals in different ways and still applying source separation-like algorithms exciting results would ensue. Particularly so in sensory perception research we saw the automatic discovery of behavior we find in our sensory systems by just using sparse coding ([6] for vision, [7] for audition). As research in sparse coding blossomed it also found applications in the field of music where we saw the use of sparse coding to perform music transcription ([3], [4]). Although there has been excitement about the intuitive features that one can find in musical signals using sparse coding, the connection to source separation has been somewhat neglected.

In this paper we present a new sparse coding algorithm which we originally developed for music analysis but also came to be very applicable for source separation. For the remainder of this paper we go through some of the research steps we have taken which lead us from musical analysis to source separation and come back full circle to an application for musical signals.

2 Learning Musical Elements

The first thing a music student learns is that music is composed out of some basic elements. To the trained ear a musical piece is not treated as one long continuous sound, but rather a mixture of notes, chords, and other well known musical constructs. Although a considerable effort in research has been expended in making computers detect these constructs it still remains an elusive process. Partly this is because the mathematical definition of a note, that a computer would require, can be a hard thing to express. In this paper we approach the problem of learning music from a somewhat evolutionary perspective. We will not explicitly require that our system looks for notes, but rather we will let it find by itself what is the best way to formalize music. Surprisingly enough we will see that notes are indeed discovered as an optimal way of representing music.

As in most audio analysis systems we will start from a spectrogram representation. We will view the magnitude of the spectrogram as a probability distribution of acoustic energy in the time-frequency plane. Viewing it as such allows us to use it to perform latent variable analysis on it using the Probabilistic Latent Semantic Analysis (PLSA) algorithm [1]. Denoting the magnitude spectrogram by $S(f,t)$ for each frequency f and time t we define the PLSA decompositions as [1]:

$$S(f,t) = \sum_{i=1}^{c} p_i W_i(f) H_i(t) \qquad (1)$$

The objective is for a given c to find the appropriate distributions W, p and H that satisfy the equality as best as possible. Before we continue on describing how this can be done, let us consider what these variables represent. We can consider $S(f,t)$ to be a 2-dimensional distribu-

[1] In this formulation we avoid the usage of probabilistic notation to describe PLSA, and rather use notation more familiar in the field of acoustics. Although the PLSA formulation presented here looks different from other sources, it is computationally the same.

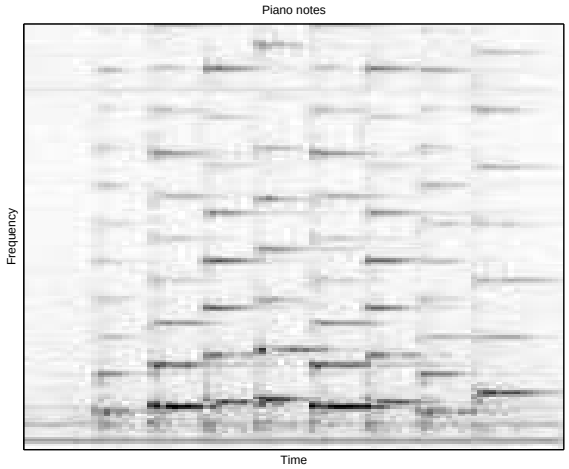

Figure 1: Input spectrogram of a piano note sequence. The particular passage contains eight note events composed out of five district notes, the sequence is: *do-re-mi-fa-re-mi-do-sol*

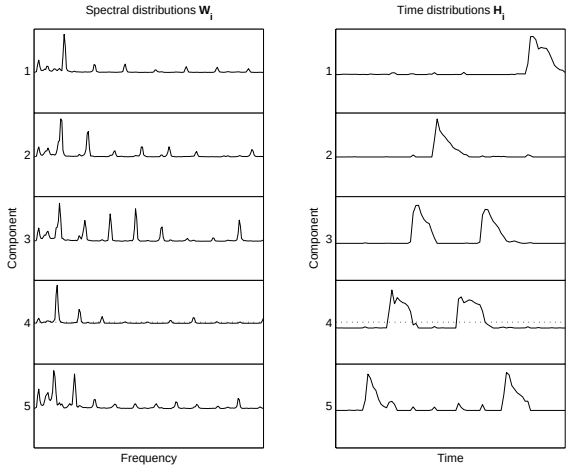

Figure 2: PLSA results after analyzing the data in figure 1. The left panel shows each W_i, and the right panel each H_i. Note that each W_i corresponds to a harmonic series template which describes each note in the input, and H_i is the corresponding note's energy envelope in time. The correspondence from component number to note is: Component 1 = *sol*, Component 2 = *fa*, Component 3 = *mi*, Component 4 = *re*, Component 5 = *do*. The order of components is arbitrary, here they are plotted in frequency order.

tion, if $c = 1$ then W_1 and H_1 will be the marginal distributions of S. If $c > 1$ then we are essentially describing a weighted mixture of marginal products that approximate S. The terms p_i are the weights of each marginal pair product. In terms of spectrogram terminology each W_i would represent a spectrum and each corresponding H_i a time envelope. So essentially this decomposition will describe a spectrogram as a set of spectra W_i modulated in time by a set of corresponding energy envelopes H_i.

The way we can determine W_i, H_i and p_i is by using the Expectation Maximization (EM) method [2]. Their estimation involves two steps, the expectation step in which we find the 'contribution' to S of each W_i and H_i, and the maximization step where we use these contributions to extract new estimates of W_i and H_i. Successive iterations lead to a solution. More specifically in our case the expectation step contributions are defined as:

$$G_i(f,t) = W_i(f)H_i(t) \tag{2}$$

And the refined estimates for each W_i and H_i are extracted from the input S weighted by the corresponding contribution in the maximization step:

$$W_i(f) = \sum_{\forall t} \frac{G_i(f,t)S(f,t)}{\sum_i G_i(f,t)} \tag{3}$$

$$H_i(t) = \sum_{\forall f} \frac{G_i(f,t)S(f,t)}{\sum_i G_i(f,t)} \tag{4}$$

Finally we need to normalize all W_i and H_i to sum to unity so that we ensure that they are true distributions. To do so we also need to derive p_i:

$$p_i = \sum_{\forall f} W_i(f) = \sum_{\forall t} H_i(t) \tag{5}$$

And subsequently use it to normalize W_i and H_i:

$$W_i(f) \leftarrow \frac{W_i(f)}{p_i} \tag{6}$$

$$H_i(t) \leftarrow \frac{H_i(t)}{p_i} \tag{7}$$

We then repeat the entire process starting from equation 2 with the newly obtained W_i and H_i and keep doing so until we observe negligible changes in them.

An additional step from standard PLSA is taken which to raise the elements of each W_i to a power $\tau \leq 1$. In the first iterations we set τ to a value around 0.8 and we progressively increase it so that by the end of training $\tau = 1$. This is a simple way to enforce that most energy from S will be distributed to W_i instead of H_i. The motivation for this will become clear later on as we examine the nature of W_i and H_i (H_i will effectively correspond to weights positioning each W_i to perform an approximation, by forcing the W_i to be 'busier' we impose more coding sparsity).

Now in order to see how these variables relate to musical spectra let us consider an example of piano music. The spectrogram of the chosen signal is shown in figure 1.

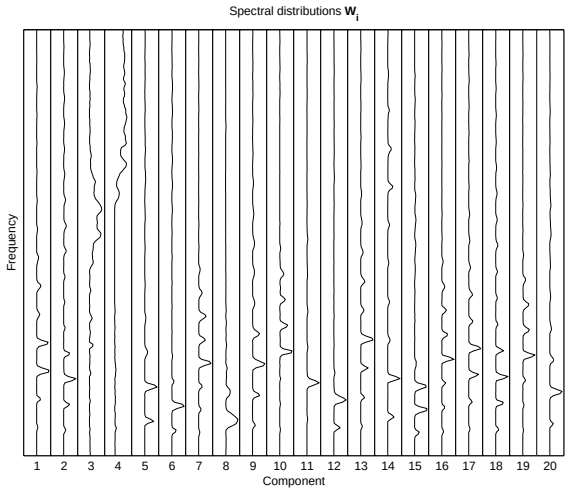

Figure 3: The extracted spectral distributions from a speech signal. Note how the distributions model various phonemes, most of them being harmonic representing vowels and a couple being high frequency wideband representing consonants.

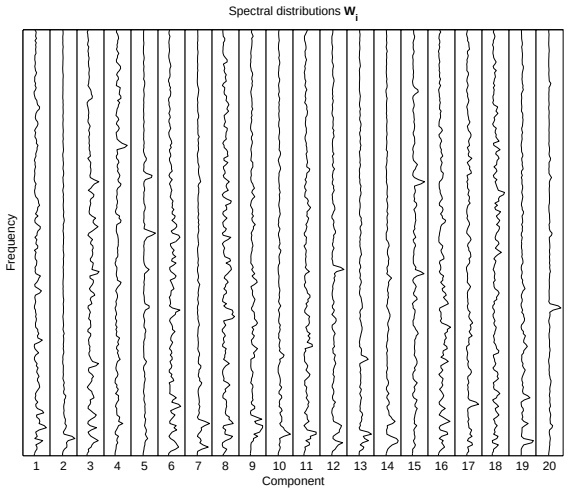

Figure 4: The extracted spectral distributions from an ambient street noise signal. Note how the spectral distributions are more wideband and less structured as compared to the speech distributions in figure 3 since they are modeling a spectrogram of a noisier nature.

Those versed in the art of spectrogram reading can make out the fact that there are eight note events, those who excel in it can tell that there are five unique notes three of them being repeated twice. More specifically the note sequence we recorded is *do-re-mi-fa-re-mi-do-sol*, the first 8 noted from Bach's Invention I in C Major (BWV 772). We will apply PLSA on this spectrogram and see how the results correlate to our knowledge of the input. We do so with $c = 5$ and derive the W_i and H_i shown in figure 2

By observation of the results we note some interesting facts. Each W_i corresponds to a (colored) harmonic sequence of a unique piano note, and each corresponding H_i shows how much energy this note had at any time. The variables p_i are the priors of the notes, essentially telling us how much each note is present. This is an interesting outcome since we have ended up extracting high-level musical information from a simple probabilistic factorization concept. Obviously the above experiment is a simple small scale example, for more elaborate musical transcription examples and insight the reader is referred to [3] (although the computational approach in that paper is different it is actually yielding qualitatively the same results as PLSA).

3 Learning General Sound Elements, and Using them for Separation

The PLSA components that we derived as notes in the preceding section need not be as coarse parts of the signal. Instead of requesting a number of components comparable to the number of notes, we could request a number much larger which would result into components that describe finer elements of the input, such as note attacks, noise transients, harmonic segments etc. Perhaps a part where this is better understood is speech. Consider the previous experiments but this time using speech instead of piano sounds as an input. The phrases we used were taken from the TIMIT database and comprised about $30sec$ of audio. The extracted W_i are shown in figure 3.

Observing the bases W_i we can see that some appear as harmonic series which correspond to vowels, and a couple appear as noise bands which correspond to consonants. In effect we discover the components that describe speech, just as we did before with music. Only this time the building elements of this domain are phonemes and not notes. We can repeat the same experiment for other types of sounds as well and we often obtain basic building blocks that fit very well to semantic descriptions we would use.

The fact that this procedure comes up with the building blocks of each type of sound it is presented, allows us to use it for separation tasks. Let us consider the case of the above speech data amidst ambient noise. The equivalent W_i from an ambient noise signal are shown in figure 4. Note how the two sets differ and encapsulate the nature of the signals they represent. The speech spectral distributions are predominantly harmonic as speech tends to be, whereas the ambient noise spectral distributions are more wideband and noisy, better describing their type of sound. Had we been confronted with a situation where there was a mixture of speech and ambient noise, it is safe

to assume that the mixture spectrogram will contain some mix of both the noise and the speech spectral distributions, each describing the presence of each sound type in the signal. This means that if we know these spectral distributions beforehand we can try to reconstruct the now unknown mixture using them. This also means that the subset of spectral distributions that describes the speech would most likely account for the speech part of the mixture, whereas the spectral distributions that describe the ambient noise will do so for the noise. We can therefore create selective reconstructions of the mixture using one spectral basis subset at a time to extract individual sound classes. We proceed with validating this suspicion.

In order to do this we need to train on models of the two sound types in advance. We have already done this and have derived the relevant spectral distributions for both speech and noise as shown in figures 3 and 4. We call these two sets of spectral distribution sets $W^{(s)}$ and $W^{(n)}$. We now obtain a spectrogram of a mixture of these two sound types. The mixture we used contained sounds that are similar in nature to what we have trained on, but not the same. We perform PLSA on this as well only this time we consolidate $W^{(s)}$ and $W^{(n)}$ into a set W_i containing both and use that as a fixed value during EM training, thereby only estimating H_i and p_i. W_i, H_i and p_i will have twice as many entries as they would have had during training for the individual sounds since they correspond to twice as many spectral distributions. Half of the elements of H_i and p_i will correspond to $W^{(s)}$ and the other half to $W^{(n)}$. We can now try to resynthesize the mixture spectrogram using only the distributions corresponding to one sound type. To do so we perform:

$$\hat{S}_{speech}(f,t) = \sum_{\forall i \text{ from } W^{(s)}}^{c} p_i W_i(f) H_i(t) \qquad (8)$$

$$\hat{S}_{noise}(f,t) = \sum_{\forall i \text{ from } W^{(n)}}^{c} p_i W_i(f) H_i(t) \qquad (9)$$

In order to go back to the time domain we invert the spectrograms $\hat{S}_{speech}$ and $\hat{S}_{noise}$ using the phase from the original mixture spectrogram. In effect what this procedure does is modulate the energy of every frequency to in the original spectrogram to look like the type of data that the selected spectral bases described. The original mixture spectrogram and $\hat{S}_{speech}$ are shown in figure 5. One can see speech with noise in the left panel, and the extracted speech spectrogram ($\hat{S}_{speech}$) in the right panel (due to its not particularly informative visual structure $\hat{S}_{noise}$ is not shown). Listening to the resulting sounds also verifies that indeed we can perform separation this way.

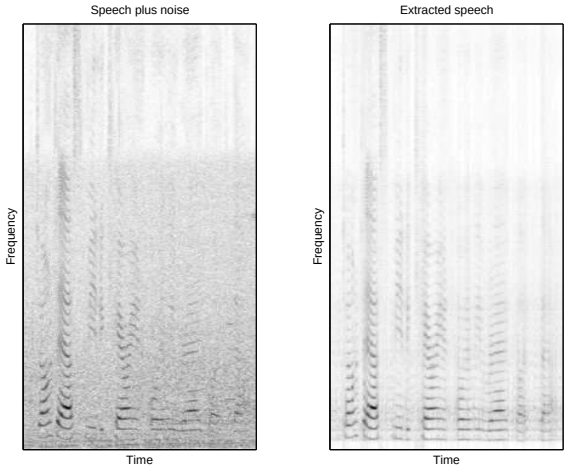

Figure 5: A speech plus noise spectrogram is shown in the left panel and the extracted speech spectrogram in the right panel. Note how the noise has been significantly suppressed and most of the extracted signal is described by a formant structure (which is what the speech spectral distributions have adapted to).

4 Back to Music

We now come full circle and return to music to apply what we now have done in a musical context. The problem we will now address is slightly different. We will be trying to separate singing voice from an accompanying piano. This time we will have an extra complication that we will only have one of the sounds in our disposal.

This is in a sense a similar problem to the previous separation case. The learned spectral distributions from the piano and voice to be very note-like and in effect describe what these two instruments were playing. Therefore in our analysis we will be learning the musical elements of the two sounds and then selectively reconstructing the piece using only the 'notes' of each one. The added complication we will have in this case is that we will be doing the analysis from a real recording which features an isolated piano part in the introduction, but no isolated vocal part from which to learn the vocal spectral distributions.

To address this peculiarity we proceed as follows. We learn the spectral distributions $W^{(p)}$ of the piano part from the musical segment in which it is isolated. We then move on to a passage where both piano and voice are present. We know that the piano presence of that segment should be adequately described by the spectral bases, but the vocal would not. Therefore we modify the learning procedure to implement this statement. We make some random spectral distributions for the voice $W^{(v)}$ and pretend that they are adequate for describing what the piano bases can not. We then go on to combine the two distribution sets as we did before on the mixture, only this time

instead of learning only H_i and p_i, we also learn half of the spectral distributions (the ones that correspond to the vocal bases, thereby leaving only the piano distributions clamped to fixed values). Doing this is very straightforward thanks to the flexibility of the EM algorithm, we essentially perform the learning rules described above, but do not evaluate them for the spectral distributions that correspond to the piano. This way we can extract a set of vocal distributions and learn how they are combined in the signal to form whatever the piano distributions cannot explain. From this point we essentially perform the same steps as before to extract the two spectrograms that correspond to the two sounds (which are shown in figure 6). Just as before listening tests verify a satisfying separation.

As expected visual examination of the piano and vocal bases reveals a note-like structure for the piano and vowels at varying pitches for voice (due to space constraints the bases are not shown, however they are very similar to the piano and vocal bases presented in previous sections).

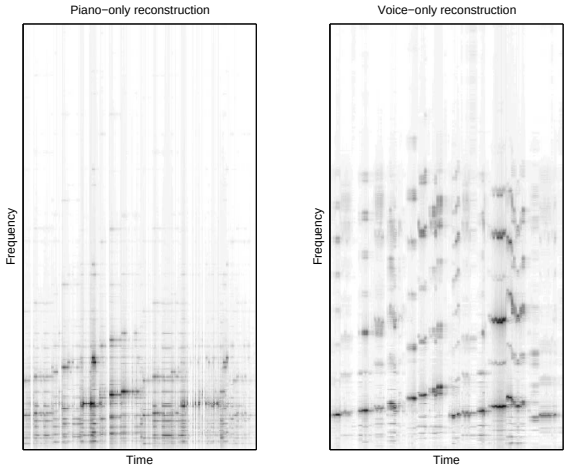

Figure 6: The extracted spectrograms of the piano (left panel) and the vocal part (right panel). Note how the piano part has more constant harmonics (since the piano has no vibrato), as opposed to the vocal spectrogram which exhibits more dramatic pitch fluctuations.

5 Conclusions

In this paper we presented a new sparse coding approach and some of its implementations for analyzing musical signals and also perform source separation under various conditions. We showed how the same process that extracts musically meaningful features from music signals can be used to extract equally important semantic representations of other types of sounds, and how these elements can be used as a basis with which to perform source separation between different sounds. Coming back to music we showed how we can apply this technique in a musical setting and also perform musical separation in a slightly different context in which training data were not available for all sound types.

The goal of this paper was to expose the reader to the connections between research in sparse coding based feature extraction and source separation and place it in a musical signal processing context. The algorithm used was only a chosen tool for sparse coding, the same results can be also obtained using a variety of other sparse coding techniques (as described in several of the referenced papers). In fact a variety of other music related tasks can be performed using this reasoning, this is still an evolving field and new and interesting papers are continuously coming out.

The author wishes to acknowledge the insightful help of Bhiksha Raj of Mitsubishi Electric Research Laboratories in formulating PLSA for audio analysis.

References

[1] Hoffman, T. "Unsupervised Learning by Probabilistic Latent Semantic Analysis", Machine Learning, vol. 42, pp. 177-196, 2001.

[2] Dempster, A.P., N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," In *Journal of Royal Statistical Society B*, vol. 39, no. 1, pp. 1-38, 1977.

[3] Smaragdis, P. and J.C. Brown, "Non-negative Matrix Factorization for Polyphonic Music Transcription", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA), pp. 177-180, October 2003.

[4] Abdallah, S.A. and M.D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra", in *Proceedings of the 5th International Conference on Music Information Retrieval* (ISMIR 2004), Barcelona, Spain, October 10-14, 2004.

[5] Hyvärinen, A. "Survey on Independent Component Analysis", available online at: http://www.cis.hut.fi/aapo/papers/NCS99web/

[6] Olshausen B.A. and D.J. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images", in *Nature*, 381: 607-609, (1996)

[7] Smaragdis, P. "Redundancy Reduction for Computational Audition, a Unifying Approach", Ph.D. dissertation, Massachusetts Institute of Technology, 2001.